A Recipe for Watermarking Diffusion Models

Abstract

Diffusion models (DMs) have demonstrated advantageous potential on generative tasks. Widespread interest exists in incorporating DMs into downstream applications, such as producing or editing photorealistic images. However, practical deployment and unprecedented power of DMs raise legal issues, including copyright protection and monitoring of generated content. In this regard, watermarking has been a proven solution for copyright protection and content monitoring, but it is underexplored in the DMs literature. Specifically, DMs generate samples from longer tracks and may have newly designed multimodal structures, necessitating the modification of conventional watermarking pipelines. To this end, we conduct comprehensive analyses and derive a recipe for efficiently watermarking state-of-the-art DMs (e.g., Stable Diffusion), via training from scratch or finetuning. Our recipe is straightforward but involves empirically ablated implementation details, providing a foundation for future research on watermarking DMs. The code is available at https://github.com/yunqing-me/WatermarkDM.